LEAD SCORING CASE STUDY SUMMARY

PROBLEM STATEMENT:

- X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOAL:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

SUMMARY:

I): READING AND UNDERSTANDING DATA:

Step 1: We have imported the required python libraries.
Step 2: Read and inspected the data.

II) EDA

Step 1: We have checked the Lead Number and Prospect ID column for any duplicates. And we chose to drop them.
Step 2: We have fixed null values by dropping the columns for which the null percentage is greater than 30%. Also we removed the columns that are not useful for our analysis.
Step 3: We have few columns with value ‘Select’ which means the leads did not choose any given option. Also, We have imputed/ removed the values accordingly.
Step 4: Performed outlier treatment with the help of boxplots.
Step 5: We have performed Analysis on Categorical and Numerical columns.
Step 6: Through Bivariate Analysis and Heat map we could observe the correlation between numerical variables.

III) DATA PREPARATION:

Step 1: In our Dataframe "Do Not Email" column we have Yes or no values we have converted them to 0/1 respectively.
Step 2: We created dummy variables for the categorical columns.
Step 3: Removed all the repeated and redundant variables

IV): TEST-TRAIN SPLIT:

Step 1: We have divided the dataset into test and train sections with a proportion of 70-30% values.

V) FEATURE RESCALING:

Step 1: We used the Min Max Scaling to scale the original numerical variables.
Step 2: We used a heatmap to check the correlations among the variables.
Step 3: We have dropped the highly correlated dummy variables.

VI) MODEL BUILDING:

Step 1: Using the Recursive Feature Elimination (RFE) approach, we proceeded with the 15 top important features.
Step 2: We have built the models using the RFE approach.
Step 3: Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
Step 4: At last, we have got the 11 most significant variables. The VIF's for these variables were also found to be good.

VII) MODEL EVALUATION:

Step 1: With our final model, we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity for both test and train sets.
Step 2: We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 86% which further solidified the model.
Step 3: Also checked if 80% cases are correctly predicted based on the converted column.
Step 4: Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.
Step 5: Calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 77.52%%; Sensitivity= 83.01%; Specificity= 74.13%.

VIII) CONCLUSION:

- Upon checking both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately near to the respective values calculated using trained set.

- Lead score is calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

- Therefore, we can consider this model as a good one.

- Features which contribute more towards the probability of a lead getting converted are:
  - i) Lead Origin_Lead Add Form
  - ii) What is your current occupation_Working Professional
  - iii) Total Time Spent on Website.